DHQ: Digital Humanities Quarterly
2016
Volume 10 Number 1

The Sentences Commentary Text Archive: Laying the Foundation for the Analysis, Use, and Reuse of a Tradition

Jeffrey Charles Witt <jcwitt_at_loyola_dot_edu>, Loyola University Maryland

Abstract

In this paper, I offer an overview of an idea for a metadata archive, called the *Sentences Commentary Text Archive*, that attempts to collect and make accessible metadata about the five century-long medieval tradition of commenting on the *Sentences* of Peter Lombard. If scaled for production, this kind of archive would enhance collaboration among editors, promote previously impossible analyses of large sections of the Sentences commentary tradition, and generally become the backbone of future applications making use of this data.

I. What is a “Sentences” Commentary?

Peter Lombard wrote the medieval book, known as the *Sentences*, during the course of his teaching career at Paris in the middle of the twelfth century. In this work, Lombard intended to bring together diverse theological opinions (*sententia*) and to harmonize them into a single concordant system. The book soon became the standard theological textbook in European theology faculties. By the thirteenth century, writing a commentary on the *Sentences* of Peter Lombard was a required exercise for anyone hoping to become a Master in the faculty of theology.[1] 1

Given that Lombard's *Sentences* covers a broad spectrum of philosophical and theological issues, it is an extremely valuable source for tracking the development of ideas. Today, we know of approximately 1,000 extant commentaries. Unfortunately, only a small number of these commentaries are available in printed editions, and an even smaller number are available in critical editions. Most remain available only in difficult to read Latin manuscripts stored away in difficult to access libraries. 2

While the current scholarly practice primarily aims to make these texts available in printed editions, this workflow suffers from a couple of problems. Today, a static print edition of these commentaries remains the gold standard. This is not always without good reason. The unchangeable nature of the text helps create stable citation practices that online resources often have a hard time replicating. Likewise, the print edition is easier to document for purposes of tenure and promotion. In the present proposal, I am not interested in disputing these realities, but rather in thinking about a workflow towards print that can enable other possibilities at the same time. One such possibility lies in making working drafts and on-going editions of texts — editions that often take decades to complete — sufficiently discoverable so as to promote collaboration. It is, however, hard for potentially interested scholars to collaborate if they are unaware of texts or editions in need of collaborators. Another important possibility is the future large-scale analysis of the tradition as a whole. The divergence of editorial practices, file formats, and the generally isolated practice of print production makes large-scale analysis of the entire corpus difficult. Often very little is known about the digital format of the post-publication file. But these post-publication files, if archived and catalogued in an accessible way, could become the bases of a new understanding of this five-century long tradition and its various developments across time and region. 3

The Function of a Metadata Archive

In this short paper, I offer an overview of a metadata archive (still in development) that, if scaled for production, could 4

support the kind of collaboration mentioned above, promote previously impossible analyses of large sections of the *Sentences* commentary tradition, and generally become the backbone of future applications making use of this data.

5 By “archive” I mean a large collection of resources produced by an identifiable tradition and organized (with intentional metadata) according to the organic relationships that constitute said tradition. While Kate Theimer has noted the propensity of digital humanists to play fast and loose with the meaning of an “archive,” she mentions one important criterion of an archive that fits my present use of the word. In attempting to describe the difference between an archive and a collection she writes:

> What defines the work of archivist, and so “an archive” in the mind of an archivist, is what materials are selected and how they are managed. Archivists select and preserve “archives” as defined in the primary definition, *which is to say aggregates of materials with an organic relationship*, rather than items that may be similar in some manner, but otherwise unrelated (emphasis mine).[2] [Theimer 2012]

6 The suggestion that an archive is a collection of materials with an *organic relationship* is what I have in mind here. This metadata archive attempts to aggregate a collection of intimately related resources that build on and reference each other over time. The goal of the metadata is to organize these resources in such a way that organic relationships can be tracked and made discoverable by future users.

7 Accordingly, this metadata is the beating heart of the archive and the basis on which a user or application can search and sort resources. As such, it also true to say that this conception of archive does depart from the perhaps more traditional notion of an archive as a collection of material objects and is instead primarily interested in a “purposeful collection of digital surrogates” [Price 2009, 22].

8 The need for such a metadata archive was first impressed on me as I worked (and continue to work) on an edition, encoded according to the standards of the Text Encoding Initiative (TEI), of a very late-fourteenth century *Sentences* commentary written by a Parisian thinker named Peter Plaoul. What I now call the *LombardPress* web publication framework http://lombardpress.org originally started as a way to progressively display the ongoing transcription of Plaoul's commentary in a user-friendly way. (See http://petrusplaoul.org.) The goal was, and still is, to allow users access to early transcriptions and to allow them to comment and collaborate, should they become interested in the text: an approach that has much in common with what is elsewhere called “interactive dynamic editing” [Pichler and Bruvik 2014, 183]. The edition began in March 2011 and has now made available over 300,000 Latin words present in four diplomatic transcriptions, totaling over 1,200,000 words.

9 As the edition developed, I noticed two things. First, this semantically encoded text is rich in interesting metadata (e.g. citations, references, structure, length, word-frequency, etc.) that, in order to be useable, simply needed to be harvested and presented. Second, as I began to detach the *LombardPress* framework from the text of the Plaoul edition in order to support the presentation of other editions (see for example http://adamwodeham.org), I noticed that there was a lot of repeated and re-useable information. The prospect of a central archive would not only allow different applications to reuse redundant information, but would also allows users and applications to recognize interesting connections across the entire tradition: connections like the reuse of a particular quotation by multiple authors or spikes in the use of key words across time.

III. Technical Specifics: The Current *Sentences Commentary Text Archive* Prototype and Workflow

10 A primary goal of the *Sentences Commentary Text Archive* is to be a coordinating service that both humans and machine applications can use. While the archive could itself include the capacity to store the actual source text of an edition, this is not its central function. Rather, as individual sections of a text begin to be edited, it is recommended that they be stored in a development repository (e.g. http://github.com or http://bitbucket.org). Subsequently, strategically selected versions should be deposited into an institutional repository. The proposed metadata archive would then

become a central clearinghouse of information about these texts that can, first and foremost, point users, developers, and publishers to the location of the encoded text, as well as to any plaintext or HTML manifestations. With these pointers in place, substantial amounts of metadata can be added to facilitate the discovery of various texts.

Figures 1 illustrates, in a highly abstract way, how this metadata archive could serve as a kind of switchboard between institutional and development repositories of texts and application uses of those texts. 11

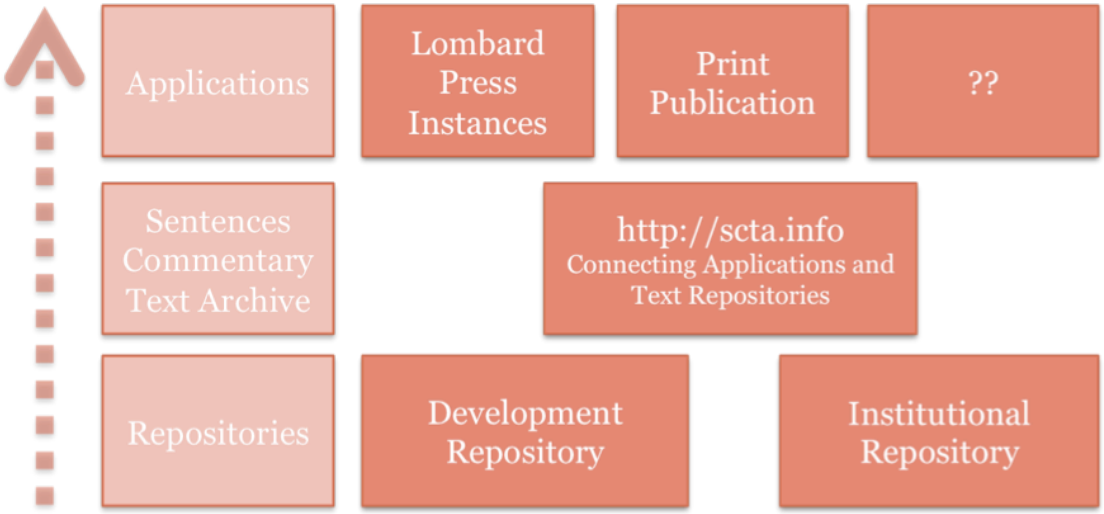

Figure 1. Illustration of metadata archive as switchboard between repositories and multiple applications

In this section, I want to provide some fairly specific technical details about the current prototype of the archive that I have built. The point of such a report is not necessarily to declare the best possible practice or even describe what the final form of the *Sentences Commentary Text Archive* should look like. Rather, the primary point is to identify one possible (and tested) route by which the idea might be implemented. Second, it is to simply demonstrate that the technical challenges are the smallest obstacles to implementation. Technically speaking such an archive is not overwhelmingly difficult to construct. The larger obstacle is finding institutional support and generating the collective will of the field to contribute its semantically encoded transcriptions (both in draft and final form) for metadata harvesting. 12

In the current prototype (an early incarnation of which can be found at http://scta.info), the archive metadata extraction begins with an XML document, called a "projectdata.xml" file that includes a list of `<item>`s. Each `<item>` in this document represents a discrete section within a given *Sentences* commentary. Generally, it is assumed that an `<item>` is a natural place for discrete file divisions to be made and thus each `<item>` would correspond to a TEI file replete with its own metadata-rich `<teiHeader>`. This kind of list of `<item>`s is usually called a “Question List” when published in academic journals. See for example [Friedman-Schabel 2001] or [Marcolino 2008]. Since these lists are the starting point of the archive, one future goal is to solicit new “question lists” encoded according to an established schema wherein the editors of each list — both the compiler of the original list and the encoder of the data — are clearly credited for their work. Our current working database includes eight such lists including over 1,300 indexed `<item>`s. There are, however, approximately 1,000 possible lists and therefore tens of thousands of question `<item>`s to be indexed. Depending on the structure of the commentary in question, these `<item>`s can be nested inside grouped sections, typically distinctions or books: a relationship that will be recorded in the extracted metadata. It is also possible that information about individual diplomatic transcriptions, including the folio numbers where this `<item>` of text can be found, might be nested within a given item. This foundational document will also include pointers to the repository where the raw TEI text of each item and any individual diplomatic transcriptions may be found. As part of the transition from prototype to implementation, an XML schema will need to be constructed to make sure each of these files adheres to a recognizable standard. 13

The standardization of these files is critical for the next step in the process of extracting metadata according to the 14

Resource Description Framework (RDF) data model.[3] An XSLT script, “rdfextraction.xsl,” is applied to a given “projectdata.xml” file.[4] The result is a large set of RDF triples in an XML format that can easily be imported into a server that can be queried using the SPARQL and RDF Query Language (SPARQL). The power of the “rdfextraction.xsl” is that it not only processes the information in the “projectdata.xml” file, but that it follows the pointers in the file to the raw TEI text, when available, and begins extracting data for each of the XML encoded transcriptions.

From the extracted information, this archive can, first and foremost, promote collaboration by simply listing for interested researchers whether or not a transcription of the particular part of text (i.e. an <item>) is “available,” “in progress,” or “not yet started.” If a user knows the exact section of a text they are interested, they could look up the text in the database and see the status of the text. This simple step would simultaneously help scholars avoid redundant work and encourage collaboration. 15

This kind of information, however, is only the tip of the iceberg. When texts are encoded according to the TEI schema — or, in the ideal scenario, according to a customized TEI schema tailored to this specific genre of text — it is possible to begin automatically harvesting all kinds of information about the text itself. For example, when <name>, <title>, <quote>, or <ref> tags are identified with unique standardized identifiers (such as @key, @ref, or @ana attributes), we can quickly harvest information about the number of mentions of distinct authors, or the use of titles, references, quotations, or other key words and technical terms or phrases. Extraction of metadata from a thoughtfully TEI encoded edition is remarkably simple, fast, and robust. Thousands of relationships can be constructed in seconds. 16

For example, in the case of the Peter Plaoul edition, the “rdfextraction.xsl” stylesheet can, with astonishing speed, run through over 650 documents and 1,200,000 words and return information about which authors were mentioned, what texts were mentioned, and what quotations were cited. 17

Though this is the ideal method of abstraction, it remains the case today that many *Sentences* commentary editions remain in print form only. Likewise, many editors continue to edit in Microsoft Word format. While the promise of such an archive would be greatly enhanced by a field that understood the power and importance of semantic encoding, there is no use in denying the realities of current practice.[5] This is not, however, an insurmountable obstacle. There are other ways to scrape information from a plain text file, and a full implementation of this archive would need to develop strategies for these common cases.[6] 18

More essential than exactly how this information is harvested is the schema according to which this metadata is catalogued. My present prototype implementation has a created a provisional RDF schema. This schema is divided into three primary classes: texts, resources (such as names, works, quotations, etc.), and properties (such as hasTranscriptions, quotes, isQuotedBy, mentions, references, etc.). Properties are further divided into three main categories that allow the archive to organize metadata according to three main data streams, publication information (pubInfo), content information (contentInfo), and linking information (linkingInfo). Figure 2 offers a representation of how these content streams can be presented to users. 19

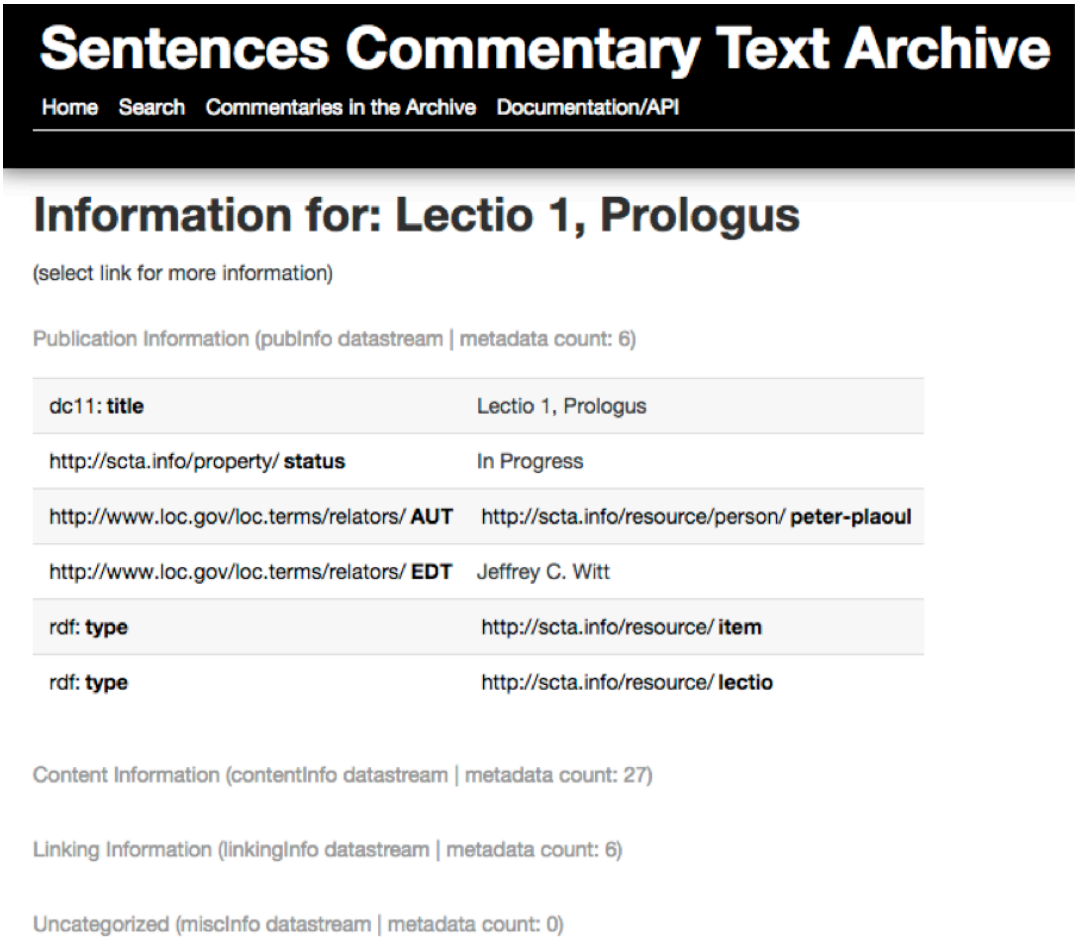

Figure 2. HTML presentation of various metadata content streams made available by the archive

Ideally this is a schema that would be developed by a team of editors and constructed in concord with the TEI customization schema. Compatibility between these two schemas makes metadata harvesting extraordinarily efficient. 20

This kind of metadata collection allows for robust search and finding possibilities. When in place, users can search for any text where a particular Bible verse is used or any text that discusses a key word such as “faith,” “baptism,” etc. Likewise, by navigating the archive they could discover networked connections such as sets of texts that quote the same authors in the same places or use the same key words in the context of the same quotations. Right now the information needed for this kind of search is scattered: scattered throughout printed critical editions and various online editions. The *Sentences Commentary Text Archive* aims to bring this information into a unified web. 21

The current prototype instantiation of the archive stores the extracted RDF triples into a Fuseki triple store, part of the Apache Jena Java framework[7] that allows for SPARQL queries to be sent over HTTP.[8] The front-end of this prototype archive is written in Ruby using the Sinatra web framework.[9] Using the RDF.rb library,[10] the Ruby frontend provides users and developers with an API that can be used to navigate the RDF data, make unique connections, and perform searches within and outside the data set. 22

While these possibilities are tantalizing in themselves, the real power of this archive is not primarily in the user interface, but in the further uses that other applications can make of this information. For this purpose, the RDF.rb library is equipped to handle content negotiation. Thus, when a user requests information, the browser will return nicely formatted HTML tables. However, when a machine requests information in other file formats, specified in HTTP header, (e.g. ttl, nt, json, rdf/xml), the requests are easily handled by the Ruby application. 23

IV. An Example Application: The *LombardPress* Framework

24 One good example of an application that can make use of the metadata in this archive is the *LombardPress* framework. As mentioned above, this application was initially developed to display the Peter Plaoul edition. Thus, initially the framework and the text were welded together rather tightly. The long-term goal for the *LombardPress* application is to increasingly rely on the archive and to be increasingly independent of the raw text. And as the *LombardPress* framework begins to display other texts, it can benefit from the shared information in the archive.

25 First, the *Sentences Commentary Text Archive* aims to include basic information about every author ever cited in any *Sentences* commentary. At the same time, it is also expected that each web publication will have its own prosopography and name index. Rather than requiring each collection of texts to maintain a separate prosopography, the *LombardPress* framework can simply run through the current texts it is assigned to display, identify which names are used (assuming the text has been edited according to agreed upon standard and names have been identified with agreed upon name IDs), and then query the *Sentences Commentary Text Archive* to automatically create a custom prosopography for the existing project. The framework can then use this custom list to generate a custom name index.

26 A second example lies in the gathering of text files to be displayed in a single *LombardPress* website instance. At present, an individual site powered by the *LombardPress* framework is populated by its own “projectdata.xml” file which includes a lot of redundant information already contained in the archive, such as the location of the text in institutional and development repositories, the formal title of the text, its status (draft or otherwise), and whether it has any corresponding diplomatic transcriptions, subdivisions, etc. Future development aims to reduce this file to a simple set of dereferenceable URLs that the user wants to display, in the order he or she wants them displayed. With the archive in place, the application can make a call to the archive and, by parsing the response, can access all the other pertinent information it needs. Likewise, if the location of the raw XML file changes (for example, if it is moved to a new repository), no updates will ever need to be made. The application will still know, through the information returned from the archive, exactly where the raw text can be found.

27 This kind of dependence would also allow for the prospect of building a *LombardPress* instance (or another application) that aims not to display an entire commentary, but rather to curate a selection of texts that are united by a common theme or subject matter. For instance, suppose someone wants to present a selection of texts by multiple authors on the topic of “baptism”. Suppose further that these various texts were edited by several different editors and stored in various different institutional repositories across the globe. Using the archive, it is possible first to search for and select all the texts that discuss “baptism”. Subsequently, an application can follow the metadata pointers to the location of the raw source text and then (using more associated metadata) display these texts chronologically.

28 Third and finally, the ability to observe connections through the metadata archive could be used to suggest connections and topic threads to readers as they consider one text. Future development on *LombardPress* aims to create a suggested readings list that identifies quotations or key words used in a paragraph, queries the archive for other *Sentences* commentaries that also use these quotations or key words, and then offers the reader links to the related texts.

29 These kinds of developments — the overall separation of the text from the display system and increasing reliance on a centralized archive — are important strides toward the long-term sustainability of both the text and any particular display system. More than likely, the web application will be a continual work in progress, which is simply the nature of modern web development. Thus, through rigorous separation of the text and the platform in which it is viewed, the text can survive and remain accessible through all the vicissitudes that come with the development of a web application.

30 Likewise, the dependency of an application like *LombardPress* on a stable archive can also help address some important challenges often targeted at digital publication.[11] The most pressing of these is the reliability and stability of URL citations. An institutional commitment to a metadata archive would provide two benefits in this regard. On the one hand, it would allow for stable URLs. On the other hand, it would give editors and developers the freedom to experiment with producing different manifestations of the text using temporary URLs that may not be intended to last forever, but

nonetheless remain valuable experiments in development.

Figure 3 provides an illustration of a comprehensive workflow from editing to web and print publication that makes use of the proposed archive. 31

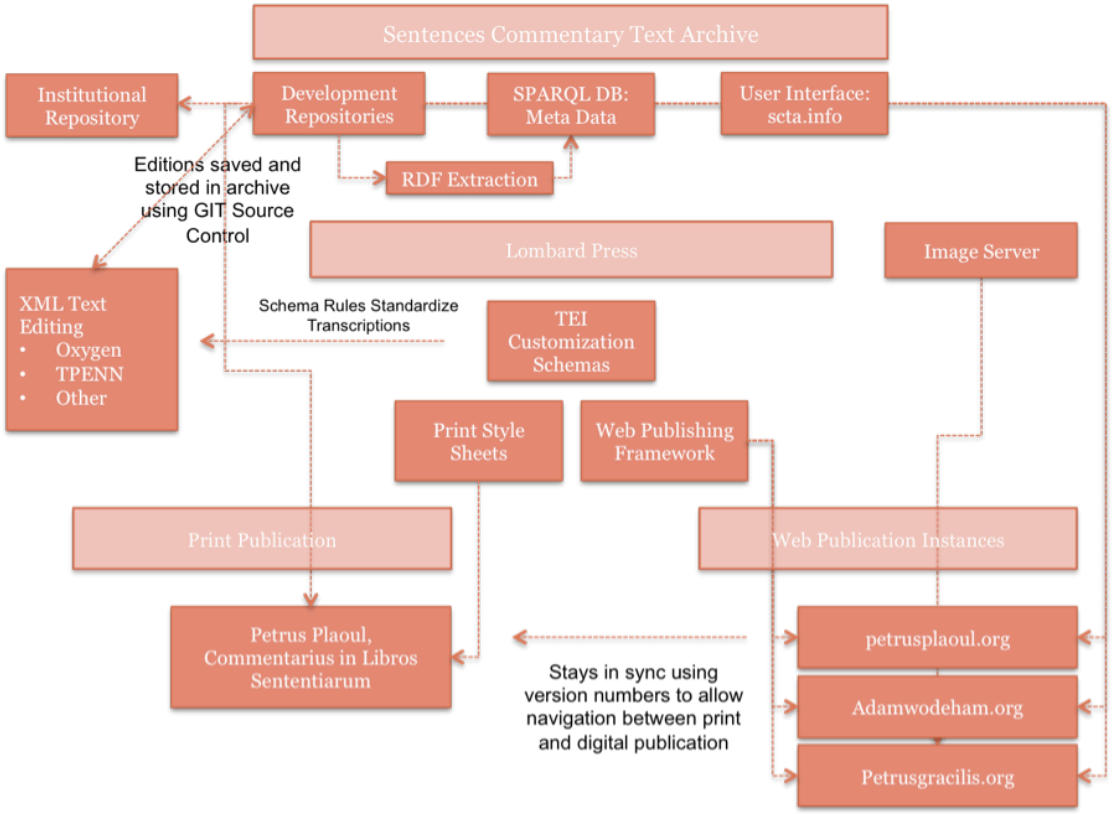

Figure 3.

V. Conclusion: Challenges and Possibility.

As noted above, the technical requirements, while possessing their own challenges, are not particularly difficult to meet. With a little bit of funding, a small team of developers could easily implement and augment the workflow described above. The biggest challenge is contribution. 32

One obvious hurdle is getting people to edit their texts according to the TEI schema or in a format easily convertible to TEI. However, I imagine that even if users are allowed to submit in any format, and different scripts are designed to harvest the metadata from the various formats, people will still be reluctant to allow metadata to be harvested from their ongoing work. 33

In response to this concern, let me conclude with a couple of pertinent facts about the above proposal. The first is that, while it is possible for the archive to be a place to actually host the XML files themselves, this is not its primary function, and it is not necessary for the text to be permanently handed over to a third party. In the end, the ideal scenario would be for the raw texts to be deposited into an editor's own institutional repository. In this case, editors and projects managers can retain control over where their files actually live. The second pertinent fact is that individual editors can still control accessibility to these texts by indicating their publication status in the `<teiHeader>`. What we are primarily talking about here is metadata abstraction. While the extraction will initially need access to the raw text, it does not need permanent access. Further, metadata tags regarding status of the text and its publication license can alert users to the availability of the text and individual editors can choose whether or not to make the full text accessible to many or to just a few. 34

In sum, two dominant principles continue to guide the present proposal. The first principle is a healthy appreciation for 35

the need to support and foster collaboration. To support this collaboration, we need to create a finding aid that gives people enough access to a text to become interested in it. The second principle is an equally healthy respect for an editor's desire to control the content he or she is working on for both development and publication purposes. I believe the *Sentences Commentary Text Archive*, as described here, is a feasible way to provide this kind of controlled access and discovery. At the same time, I believe it can be the foundation of a long-term resource that would enable future analysis of the entire *Sentences* commentary tradition that, at present, still remains out of reach.

Notes

[1] For Peter Lombard's text and structure see, in Latin, [Lombardus 1971-1981] and, in English, [Lombard 2007-2010]. For a discussion of Lombard's *Sentences* and its legacy see [Rosemann 2007, 41–51].

[2] For a further discussion of collection naming, see [Price 2009]

[3] See http://www.w3.org/RDF/ and http://www.w3.org/TR/2014/REC-rdf-schema-20140225/ for a helpful overview of the RDF data model.

[4] This process follows a general trend in humanities computing. A report from the Orlando Project (OP) writes: “Research projects and content providers in the humanities such as libraries and museums are increasingly incorporating semantic web technologies. As part of this process, many ontologies initially developed for other contexts are being translated into semantic-web-ready forms to enable the leveraging of existing metadata in a semantic web context. XML schemas in particular are targeted for such translation, and production of RDF based on existing XML markup is increasing, with the W3C offering a conversion tool meant to facilitate such translations” [Simpson-Brown 2013, 194].

[5] Despite this reality, in a recent article, Alois Pichler and Tone Merete Bruvik remind us that encoding and presentation are distinct acts, and that, when blurred, “it often happens that the first set of procedures, transcription, is biased by the second, presentation” [Pichler and Bruvik 2014]. It is hoped that the *Sentences Commentary Text Archive* will become another tool that will further encourage the separation between encoding and presentation.

[6] As an alternative to the “standard approach” to extraction through the use of XSLT and XPath queries, the Orlando Project uses a Python script to target specific regular expressions; see [Simpson-Brown 2013]. This approach provides some important pathways for including information in the archive from commentaries that are not encoded in the preferred TEI format. It may also prove a useful strategy for targeting printed texts that can be quickly and reliably OCR'd.

[7] See https://jena.apache.org.

[8] See https://jena.apache.org/documentation/serving_data/index.html.

[9] See http://www.sinatrarb.com/.

[10] See http://ruby-rdf.github.io/.

[11] See Pichler and Bruvik who note the stability of the print text as one of its traditional strengths over digital texts [Pichler and Bruvik 2014, 182].

Works Cited

Friedman-Schabel 2001 Friedman, R. L. and Schabel, C. “Francis of Marchia’s Commentary on the *Sentences*: Question List and State of Research.” *Mediaeval Studies* 63 (2001): 31–106.

Lombard 2007-2010 Peter Lombard. *The Sentences*. G. Silano (trans). Toronto, PIMS (2007–2010).

Lombardus 1971-1981 Petrus Lombardus. *Magistri Petri Lombardi Parisiensis episcopi Sententiae in IV libris distinctae,*, 2 vols. Grottaferrata (1971-1981).

Marcolino 2008 Marcolino, V. “Zum Abhängigkeitsverhältnis Der Sentenzenkommentare Der Augustinertheologen Petrus Gracilis († N. 1393) Und Iohannes von Basel († 1392).” *Analecta Augustiniana* 71 (2008): 493–529.

Pichler and Bruvik 2014 Pichler, A. and Bruvik, T. M. “Digital Critical Editing: Separating Encoding from Presentation.” In D. Apollon, C. Bélisle, P. Régnier (eds), *Digital Critical Editions*. University of Illinois Press, Urbana, IL (2014).

Price 2009 Price, K. M. “Edition, Project, Database, Archive, Thematic Research Collection: What's in a Name?”" *Digital*

Humanities Quarterly 3:3 (2009).

Rosemann 2007 Rosemann, P. W. *The Story of a Great Medieval Book.* Toronto (2007).

Simpson-Brown 2013 Simpson, J. and Brown, S. “From XML to RDF in the Orlando Project.” *Culture and Computing* 13 (2013): 194–5.

Theimer 2012 Theimer, K. “Archives in Context and as Context.” *Journal of Digital Humanities* 1:2 (2012).